AutoBayes – Automatic Synthesis of Data Analysis Programs

Bernd Fischer
Automated Software Engineering Group
NASA Ames Research Center
fisch@email.arc.nasa.gov

Data Analysis for AR

(courtesy L. Pedersen)

Typical Data Analysis Application

Given:

- instrument data (2D pixel intensities, 3D reconstruction)
- model information / assumptions (number and shape of rocks, overall structure of scenery)

Data Analysis for AR

(courtesy L. Pedersen)

Typical Data Analysis Application

Given:

- instrument data (2D pixel intensities, 3D reconstruction)
- model information / assumptions (number and shape of rocks, overall structure of scenery)

Wanted:

- rock (i.e., model-specific program that identifies rock)

Data Analysis for AR

(courtesy L. Pedersen)

Typical Data Analysis Application

Given:

- instrument data (2D pixel intensities, 3D reconstruction)
- model information / assumptions (number and shape of rocks, overall structure of scenery)

Wanted:

- rock (i.e., model-specific program that identifies rock)

Problems:

- high implementation efforts
- high reliability requirements
- model assumptions change

AutoBayes

```
model rock as "L. Pedersen's Rock Finder".
type xyz = {nat i, j; double x, y, z}.
const nat nx as "image size (x)".
const nat ny as "image size (y)".
const nat nrocks as "number of rocks".
nat c(1..nx,1..ny) as "rock id"
  where c(_,_) <= nrocks.
double mu(0..nrocks), sigma(0..nrocks)
  where 0 < sigma(_).
double mse as "mean square error".
double a, b, c as "plane parameters".
double r(1..nrocks),  h(1..nrocks),
      dx(1..nrocks), dy(1..nrocks)
  as 'sphere parameters'.
data nat pix(1..nx,1..ny) as "2D-image".
data xyz pos(1..nx*ny)    as "3D-reconstruction".
pix(I,J) ~ gauss(mu(c(I,J)), sigma(c(I,J))).
pos(K)@z ~ c(pos(K)@i,pos(K)@j) cases
  [0->gauss(c*(1-pos(K)@x/a,pos(K)@y/b),mse),
   C->gauss(h(C)**2*
            (1-((pos(K)@x-dx(C))**2 +
                (pos(K)@y-dy(C))**2)/r(C)**2),
            mse)
  ].
max pr({pix,pos}|{mu,sigma,a,b,c,r,h,dx,dy,mse})
for {mu,sigma,a,b,c,r,h,dx,dy,mse}.
```

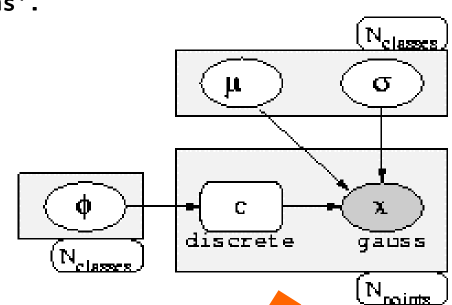
- fully automatic, end-to-end program synthesis system
- high-level, engineering-style notation
  - fully declarative
  - domain-independent
- high leverage: 1:10 – 1:30
- fast turnaround: ~150 loc/second
- code *and* documentation

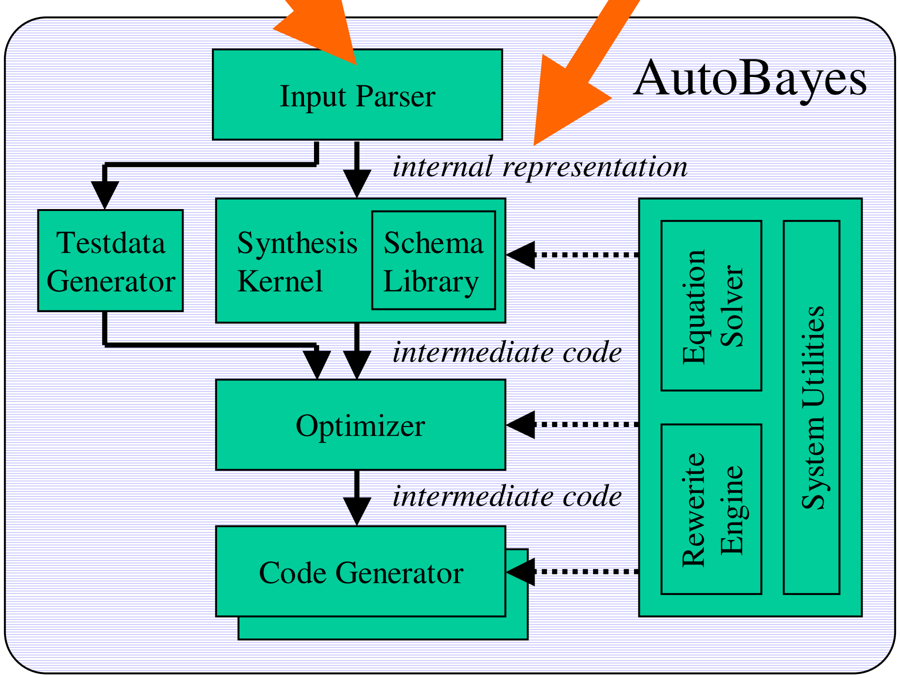
AutoBayes – System Architecture

```
model mog as 'Mixture of Gaussians'.

const nat n_points.
  where 0 < n_points.
const nat n_classes := 3.
  where n_classes << n_points.
...
double mu(0..n_classes-1).
double sigma(0..n_classes-1).
  where 0 < sigma(_).
...
data double x(0..n_points-1).
x(I) ~ gauss(mu(c(I)), sigma(c(I))).

max pr(x|{rho,mu,sigma})
for {rho,mu,sigma}.
```

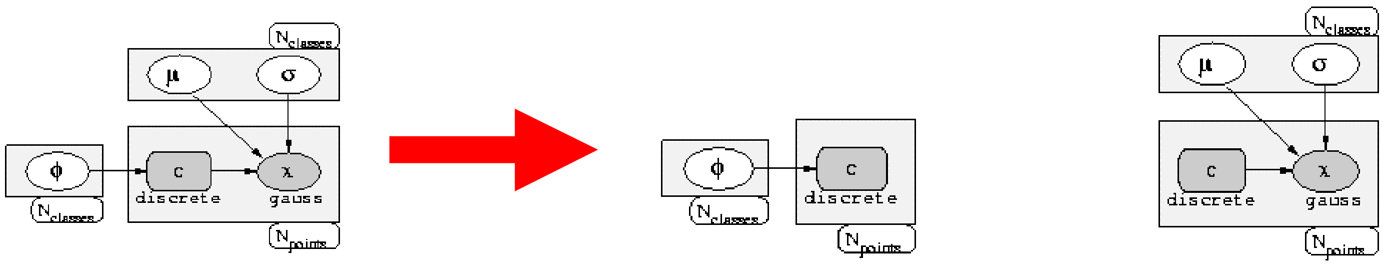




- Graphical Models
- Schema library
- Symbolic subsystem
  - rewrite engine: AC, assumptions
  - simplification
  - symbolic differentiation
  - abstract interpretation
  - (polynomial) equation solver
- Procedural intermediate language
- Multiple backends
  - C/C++ based: Octave, Matlab
  - Modula2
  - ⇒ customization (MDS, K9) possible
- Multiple programs
- ~ 50kLoC Prolog

AutoBayes – Schema Library

- Schema ≈ theorem of domain theory (cf. J. Pearl's book)

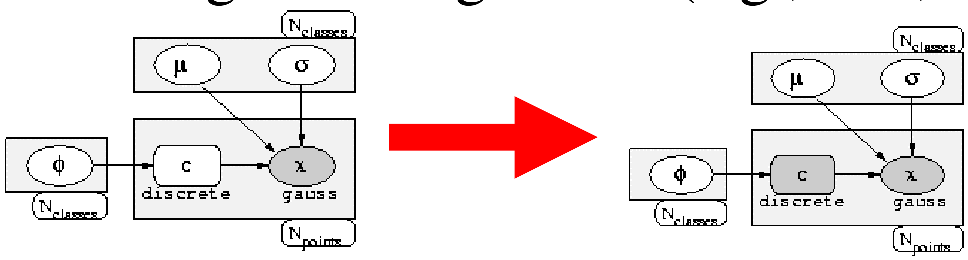



```
+ max pr({c,x}|{phi,mu,sigma}
  for {phi,mu,sigma})
```

```
+ max pr(c|phi)
  for phi
```

```
+ max pr(x|{c,mu,sigma})
  for {mu,sigma}
```

- Schema ≈ generic algorithm (e.g., EM, k-Means)



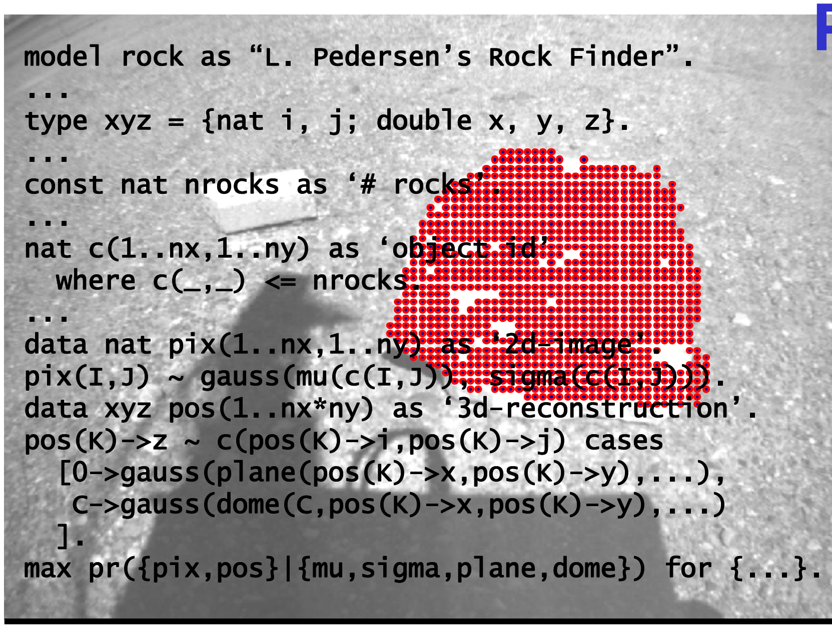

```
for i=0;i<N;I++ do {
  for j=0;j<M;j++ do {
    q[i,j]:=0.0
  }
};
converge(phi,mu,sigma){
  <max pr({c,x}|{phi,mu,sigma}
   for {phi,mu,sigma}>
};
...
```

```
+ max pr(x|{phi,mu,sigma}
  for {phi,mu,sigma})
```

```
+ max pr({c,x}|{phi,mu,sigma})
  for {phi,mu,sigma}
```

⇒ generic code fragment + applicability conditions

Automatic Synthesis of Data Analysis Programs



```
model rock as "L. Pedersen's Rock Finder".
...
type xyz = {nat i, j; double x, y, z}.
...
const nat nrocks as '# rocks'.
...
nat c(1..nx,1..ny) as 'object id'
  where c(_,_) <= nrocks.
...
data nat pix(1..nx,1..ny) as '2d-image'.
pix(I,J) ~ gauss(mu(c(I,J)), sigma(c(I,J))).
data xyz pos(1..nx*ny) as '3d-reconstruction'.
pos(K)->z ~ c(pos(K)->i,pos(K)->j) cases
   [0->gauss(plane(pos(K)->x,pos(K)->y),...),
    C->gauss(dome(C,pos(K)->x,pos(K)->y),...)
   ].
max pr({pix,pos}|{mu,sigma,plane,dome}) for {...}.
```

Objectives:

- *fully automatic generation of data analysis programs from high-level specifications (i.e., statistical models)*
- *generation of software and documentation*
- *use of optimized data structures*
- *fully automatic certification of safety properties*

Benefits:

- *reduced coding/testing efforts, increased confidence*
- *rapid turnaround of software revisions*
- *rapid exploration of design space*

Principal Investigator:

Bernd Fischer, USRA/RIACS
NASA Ames Research Center
+1 650 604 2977
fisch@email.arc.nasa.gov

Participating Organizations:

ASE Group, NASA Ames Research Center
(Fischer, Schumann, Whittle)
CMU (Gray)
HIIT (Buntine)

Achievements:

- *AutoBayes prototype synthesis system (TRL ~4)*
  - *clustering*
  - *changepoint detection*
  - *image processing*
  - *state estimation*
- *first applications to NASA data sets*
- *synthesized state-of-the-art algorithm (multinomial PCA, NIPS'99)*
- *certification of array-bounds safety policy (ITSR-funded)*
- *approach documented in conference & journal papers*